# Model-based Clustering with Dissimilarities: A Bayesian Approach [1]

Man-Suk Oh
Ewha University, Seoul

Adrian Raftery
University of Washington, Seattle

Technical Report no. 441
Department of Statistics
University of Washington

December 16, 2003

[1] Man-Suk Oh is Associate Professor of Statistics, Department of Statistics, Ewha Womens University, Seoul 120-750, Korea. Email: msoh@mm.ewha.ac.kr. Adrian E. Raftery is Professor of Statistics and Sociology, Department of Statistics, University of Washington, Box 354322, Seattle WA 98195-4322. Email: raftery@stat.washington.edu; Web: www.stat.washington.edu/raftery. Oh's research was supported by research funds from KOSEF R04-2002-00-0046-0, and Raftery's research was supported by NIH Grant 1R01CA094212-01 and ONR Grant N00014-01-10745.

# Report Documentation Page

*Form Approved*
*OMB No. 0704-0188*

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

| | | |
|---|---|---|
| 1. REPORT DATE<br>**16 DEC 2003** | 2. REPORT TYPE | 3. DATES COVERED<br>**00-12-2003 to 00-12-2003** |

| | |
|---|---|
| 4. TITLE AND SUBTITLE<br>**Model-based Clustering with Dissimilarities: A Bayesian Approach** | 5a. CONTRACT NUMBER |
| | 5b. GRANT NUMBER |
| | 5c. PROGRAM ELEMENT NUMBER |
| 6. AUTHOR(S) | 5d. PROJECT NUMBER |
| | 5e. TASK NUMBER |
| | 5f. WORK UNIT NUMBER |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)<br>**University of Washington,Department of Statistics,Box 354322,Seattle,WA,98195-4322** | 8. PERFORMING ORGANIZATION REPORT NUMBER |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | 10. SPONSOR/MONITOR’S ACRONYM(S) |
| | 11. SPONSOR/MONITOR’S REPORT NUMBER(S) |

12. DISTRIBUTION/AVAILABILITY STATEMENT
**Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

| | | | | | |
|---|---|---|---|---|---|
| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
| a. REPORT<br>**unclassified** | b. ABSTRACT<br>**unclassified** | c. THIS PAGE<br>**unclassified** | | **30** | |

## Abstract

A Bayesian model-based clustering method is proposed for clustering objects on the basis of dissimilarites. This combines two basic ideas. The first is that the objects have latent positions in a Euclidean space, and that the observed dissimilarities are measurements of the Euclidean distances with error. The second idea is that the latent positions are generated from a mixture of multivariate normal distributions, each one corresponding to a cluster. We estimate the resulting model in a Bayesian way using Markov chain Monte Carlo. The method carries out multidimensional scaling and model-based clustering simultaneously, and yields good object configurations and good clustering results with reasonable measures of clustering uncertainties. In the examples we studied, the clustering results based on low-dimensional configurations were almost as good as those based on high-dimensional ones. Thus the method can be used as a tool for dimension reduction when clustering high-dimensional objects, which may be useful especially for visual inspection of clusters.

We also propose a Bayesian criterion for choosing the dimension of the object configuration and the number of clusters simultaneously. This is easy to compute and works reasonably well in simulations and real examples.

*Key Words*: Cluster Analysis, Mixture Model, Markov chain Monte Carlo, Model selection.

# Contents

# List of Tables

# List of Figures

# 1 Introduction

Cluster analysis is the automatic grouping of objects into groups on the basis of numerical data consisting of measures either of properties of the objects, or of the dissimilarities between them. It was developed initially in the 1950s (e.g. Sneath 1957; Sokal and Michener 1958), and the early development was driven by problems of biological taxonomy and market segmentation. More recently, clustering has attracted a great deal of attention as a useful tool for grouping genes and samples in DNA microarray experiments, clustering documents on the World Wide Web and in other text databases, and grouping pixels in medical images so as to identify features of clinical interest.

For much of the past half-century, the majority of cluster analysis done in practice have used heuristic methods based on dissimilarities between objects. These include hierarchical agglomerative clustering using various between-cluster dissimilarity measures such as smallest dissimilarity (single link), average dissimilarity (average link) or maximum dissimilarity (complete link), the $k$ means algorithm (MacQueen 1967), and Self-Organizing Maps (Kohonen 2001). These methods are relatively easy to apply and often give good results. However, they are not based on standard principles of statistical inference, they do not take account of measurement error in the dissimilarities, they do not provide an assessment of clustering uncertainties, and they do not provide a statistically based method for choosing the number of clusters.

Model-based clustering is a framework for putting cluster analysis on a principled statistical footing; for reviews see McLachlan and Peel (2000) and Fraley and Raftery (2002). It is based on probability models in which objects are assumed to follow a finite mixture of probability distributions such that each component distribution represents a cluster. The model-based approach has several advantages over heuristic clustering methods. First, it clusters objects and estimates component parameters simultaneously, avoiding well-known biases that exist when they are done separately. Second, it provides clustering uncertainties which is important especially for objects close to cluster boundaries. Third, the problems of determining the number of components and the component probability distributions can be recast as statistical model selection problems, for which principled solutions exist. Unlike the previously mentioned heuristic clustering algorithms, however, model-based clustering requires object coordinates rather than dissimilarities between objects as an input. Thus, despite the important advantages of model-based clustering, it can be used only when object coordinates are given, and not when dissimilarities are provided.

In many practical applications in market research, psychology, sociology, environmental research, genomics, and information retrieval for the Web and other document databases, data consist of similarity or dissimilarity measures on each pair of objects (Young, 1987; Schutze and Silverstein, 1997; Tibshirani et al. 1999; Buttenfield, 2002; Condon et al., 2002; Courrieu, 2001; Elvevag and Storms, 2002; Priem et al., 2002; Welchew et al., 2002; Ren and Frymier, 2003). Examples of such data include the co-purchase of items in a market, disagreements between votes made by pairs of politicians, the execution of links between pairs of web pages, the existence or intensity of social relationships between pairs of families, and the overlap of university applications by high school graduates.

Even when object coordinates are given, visual display of clusters in low dimensional space is often desired since it may provide useful information about the relationships between the clusters and the underlying data generation process (Hedenfalk et al, 2002; Yin, 2002; Nikkila, 2002). One way to reduce the dimensionality of objects for visual display in lower dimensional space is multidimensional scaling (MDS). In MDS, objects are placed in a Euclidean space while preserving the distance between objects in the space as well as possible.

There are many MDS techniques in the literature. Recently Oh and Raftery (2001) proposed a Bayesian MDS (BMDS) method using Markov chain Monte Carlo. This provided good estimates of the object configuration in the cases studied, as well as a Bayesian criterion for choosing the object dimension. However, they did not consider clustering and hence clustering has to be done separately with the estimated object configuration from MDS.

In this paper, we develop a model-based clustering method for dissimilarity data. We assume that an observed dissimilarity measure is equal to the Euclidean distance between the objects plus a normal measurement error. We model the unobserved object configuration as a realization of a mixture of multivariate normal distributions, each one of which corresponds to a different cluster. We carry out Bayesian inference for the resulting hierarchical model using Markov chain Monte Carlo (MCMC). The resulting method combines MDS and model-based clustering in a coherent framework.

There are three sources of uncertainty in model-based clustering with dissimilarites: (a) measurement errors in the dissimilarities; (b) error in estimating the object configuration; and (c) clustering uncertainty. Heuristic clustering methods that cluster directly from dissimilarities, such as hierarchical agglomerative clustering and self-organizing maps, do not take account of sources (a) and (c), while source (b) does not arise there. As an alternative, one may consider a two-stage procedure which estimates the object configuration in the first stage, using a heuristic MDS method or BMDS, and carries out model-based clustering in the second stage. Sequential application of heuristic MDS and model-based clustering takes account of source (c) but not of sources (a) and (b). Sequential application of Bayesian MDS and model-based clustering considers sources (a) and (c), but separately rather than together; it does not consider source (b). In contrast, our approach accounts for all three sources of uncertainty simultaneously. Simultaneous estimation of the errors is important, because errors in the dissimilarity measures and/or the estimated configuration can affect the clustering and the clustering uncertainties, as we will show by example.

Other important issues are the choice of the number of clusters and of the dimension of the objects. Oh and Raftery (2001) proposed an easily computed Bayesian criterion called MDSIC for choosing object dimension. We extend this to determine the number of clusters as well. The resulting criterion can be computed easily from MCMC output.

Section 2 describes our model for dissimilarities, the mixture model for the object configuration, and the prior distributions we use. Section 3 describes Bayesian estimation for this model using MCMC. The Bayesian criterion for choosing the dimension and the number of clusters is given in Section 4, while the method is applied to several simulated and real data sets in Section 5. A summary and discussions are given in Section 6.

# 2 Model for Clustering with Dissimilarities

Let $\delta_{ij}$ denote the dissimilarity measure between objects $i$ and $j$, which is assumed to be functionally related to $p$ unobserved attributes of the objects. Let $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$ denote an unobserved vector representing the values of the attributes possessed by object $i$.

As in Oh and Raftery (2001), we model the true dissimilarity measure $\delta_{ij}$ as the distance between objects $i$ and $j$ in a Euclidean space, i.e., $\delta_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$. In practical situations, the true dissimilarity measure can be different from Euclidean distance and there can be measurement error in observations. We therefore assume that the observed dissimilarity measure, $d_{ij}$, is equal to the true measure, $\delta_{ij}$, plus a Gaussian error. In addition, since dissimilarity measures are typically given as positive values we restrict the observed dissimilarity to be positive. Thus, given the Euclidean distance $\delta_{ij}$, the observed dissimilarity measure $d_{ij}$ is assumed to follow the truncated normal distribution

$$d_{ij} \sim N(\delta_{ij}, \sigma^2)\ I(d_{ij} > 0), \quad i \neq j, i, j = 1, ..., n. \tag{1}$$

Note that $d_{ij}$ is related to $\mathbf{X} = \{\mathbf{x}_i\}$, called the object configuration, only through $\delta_{ij}$. To represent clustering, we assume that the object configuration is a sample from a mixture of multivariate normal distributions,

$$\mathbf{x}_i \sim \sum_{j=1}^{G} \varepsilon_j \; N(\mu_j, T_j), \tag{2}$$

where each component normal distribution represents a cluster.

We use the following priors for the model parameters:

$$\begin{aligned} \sigma^2 &\sim IG(a, b), \\ (\varepsilon_1, \ldots \varepsilon_g) &\sim \text{Dirichlet}(1, ..., 1), \\ \mu_j &\sim N(\mu_{j0}, T_j), \\ T_j &\sim IW(\alpha, B_j), \end{aligned} \tag{3}$$

where $IG(a, b)$ is the inverse Gamma distribution with mode $b/(a+1)$ and IW is the inverse Wishart distribution.

One may use a more parsimonious covariance structure for $T_j$ than the above unconstrained one. For instance, one may restrict $T_j$ to be a diagonal matrix, or let $T_1 = \cdots = T_G$, or use some other parsimonious covariance model such as those commonly used in model-based clustering (Banfield and Raftery 1993; Fraley and Raftery 2002). In that case, the priors need to be modified accordingly.

# 3 Posterior Inference

## 3.1 Markov chain Monte Carlo

It is well known that inference for mixture models can be simplified with latent variables which indicate the group memberships of objects. We define latent variables $K_i$ such that $P(K_i = j) = \varepsilon_j$ and $\mathbf{x}_i$ belongs to class $j$ if $K_i = j$, so that

$$\mathbf{x}_i | K_i = j \quad \sim \quad N(\mu_j, T_j).$$

From the prior and the model, the full conditional posterior distributions (densities) given all the other unknowns are given as

$$\begin{aligned} \pi(\mathbf{x}_i | K_i = j, \text{ others }) \;\propto\;& exp[-1/2(\mathbf{x}_i - \mu_j)' T_j^{-1} (\mathbf{x}_i - \mu_j) \qquad (4) \\ & -\frac{1}{\sigma^2} \sum_{j \neq i, j=1}^{n} (\delta_{ij} - d_{ij})^2] \prod_{j \neq i, j=1}^{n} \Phi(\delta_{ij}/\sigma) \qquad (5) \end{aligned}$$

$$\begin{aligned}
\pi(\sigma^2 | \text{ others }) &\propto (\sigma^2)^{-(m/2+a+1)} \exp\left[-\frac{1}{\sigma^2}(SSR/2+b) - \sum_{i>j} \log\ \Phi\left(\frac{\delta_{ij}}{\sigma}\right)\right] \\
(\varepsilon_1, .., \varepsilon_g | \text{others}) &\sim \text{Dirichlet}(n_1+1, ..., n_g+1), && (6) \\
(\mu_j | \text{ others}) &\sim N(\frac{n_j \bar{x}_j + \mu_{j0}}{n_j+1}, \frac{T_j}{n_j+1}), && (7) \\
(T_j | \text{ others}) &\sim IW(\alpha + n_j/2, B_j + S_j/2), && (8) \\
P(K_i = j | \text{ others}) &= \varepsilon_j \phi(x_i; \mu_j, T_j) / \sum_{k=1}^{g} \varepsilon_k \phi(x_i; \mu_k, T_k), && (9)
\end{aligned}$$

where $\phi$ and $\Phi$ are respectively pdf and cdf of the standard normal distribution, $SSR = \sum_{i=1}^{n} \sum_{j=1}^{i-1} (\delta_{ij} - d_{ij})^2$,

$$\begin{aligned}
n_j &= \sum_{i=1}^{j} I(K_i = j), && (10) \\
S_j &= \sum_{i=1}^{n} (x_i - \mu_j)(x_i - \mu_j)' I(K_i = j), && (11)
\end{aligned}$$

and $I$ is the indicator function.

Iterative generation of the unknown parameters from their full conditional distributions for a sufficiently long time yields samples of the parameters from the joint posterior distribution, and posterior inference can be done by using the samples. When a simpler covariance structure is used for $T_j$ with appropriate priors, the algorithm can be easily modified since only the generation of $T_j$ needs to be changed.

Generation of samples from the full conditional distributions of the parameters $\{\varepsilon_j, \mu_j, T_j\}$ is straightforward since the full conditional posterior distributions all have convenient forms. However, the full conditional posterior distributions of $\mathbf{x}_i$ and $\sigma^2$ do not have closed forms, and so we apply the Metropolis-Hastings algorithm (Hastings, 1970) to generate samples of $\mathbf{x}_i$ and $\sigma^2$. Oh and Raftery (2001) suggested an easy random walk Metropolis-Hastings algorithm for generating samples of $\mathbf{x}_i$ and $\sigma^2$ when $G$=1. Given the latent indicator variable $K_i$, $\mathbf{x}_i$ follows a one-component multivariate normal distribution. Thus, we can easily modify the algorithm of Oh and Raftery (2001) for generating $\mathbf{x}_i$ from its full conditional posterior distribution in the mixture model. Given $\mathbf{X}$, the distribution of $\sigma^2$ does not depend on the mixture model parameters, so that the generation of $\sigma^2$ is the same in the mixture model as in the one-component model.

To initialize the MCMC algorithm, we first run Oh and Raftery's (2001) BMDS as a preliminary run to obtain an initial guess for $\mathbf{X}$, and then run MCLUST with this initial guess.

## 3.2 Non-Identifiability

Euclidean distance is invariant under translation, rotation, and reflection of objects. Thus, the dissimilarity data provide information only about the relative locations of $\mathbf{X}$. In a Bayesian context, $\mathbf{X}$ is identified, strictly speaking, but the absolute location and orientation of $\mathbf{X}$ are defined only by the prior distribution, and in practice are very weakly identified. As a result, the relative positions of the $\mathbf{x}_i$ may have a tight posterior distribution, but their absolute positions will typically have a dispersed posterior distribution.

To get around this problem of weak identification, we use a Procrustean similarity transformation (Borg and Groenen 1998, Ch.12) which proceeds as follows: (i) Obtain an estimate, $\mathbf{X}^*$, of $\mathbf{X}$ from a preliminary run, for example the MLE or the posterior mode. (ii) Transform the sample of $\mathbf{X}$ at each iteration of MCMC so that coordinates of $\mathbf{X}$ are as close as possible to the corresponding coordinates of $\mathbf{X}^*$, where the transformation is restricted to be a composition of some or all of a translation, a rotation, and a reflection. See the Appendix for more details. Since the transformation does not change the Euclidean distances between pairs of $\mathbf{x}_i$'s, it does not change the likelihood but it approximately fixes the location and orientation of samples of $\mathbf{X}$ so that $\mathbf{X}$ itself can be stably estimated.

There is another non-identifiability problem. A mixture of density functions is invariant under arbitrary permutation of component labels. Thus, the posterior density function would be invariant under arbitrary permutation of component labels unless strong prior information is used (Stephens 2000). This may cause label switching during the MCMC iterations, hence typical averages of MCMC samples of the parameters may yield unreasonable estimates of the mixture parameters. To avoid this problem, we adopt the relabeling procedure suggested by Celeux et al. (2000) at each iteration of MCMC. See the Appendix for details.

By using the two postprocessing steps, Procrustean transformation and relabelling, we obtain stable samples of the unknown parameters from which posterior estimates can be computed.

# 4 A Bayesian Selection Criterion for Configuration of Dimension and the Number of Clusters

Posterior inference as described in the previous section presumed that the dimension, $p$, of the object configuration, and the number of clusters, $G$, are given. These are typically unknown, however, and we now propose a statistical method for choosing $p$ and $G$. Oh and Raftery (2001) suggested a dimension selection criterion for MDS, called MDSIC, which works well with Euclidean distance measures with small or moderate error size. In this section, we extend MDSIC and propose a new Bayesian selection criterion, named MIC, for choosing both $p$ and $G$ simultaneously.

We view the overall goal of our analysis as being to choose the best object configuration across the dimension $p$ and the number of clusters $G$. We therefore base our model selection criteria on $\pi(\mathbf{X}_{pG}, p, G|D)$, the posterior density function of $\mathbf{X}, p, G$, given data $D$ at $\mathbf{X} = \mathbf{X}_{pG}$, where $\mathbf{X}_{pG}$ is the best object configuration given $p$ and $G$.

Note that

$$\pi(\mathbf{X}_{pG}, p, G|D) = c \cdot f(D|\mathbf{X}_{pG}, p, G)\pi(\mathbf{X}_{pG}, p, G),$$

where $c$ is a constant, $f(D|\mathbf{X}_{pG}, p, G) = \int f(D|\mathbf{X}_{pG}, p, G, \sigma^2)\pi(\sigma^2)d\sigma^2$ and $\pi(\mathbf{X}_{pG}, p, G) = \int \pi(\mathbf{X}_{pG}, p, G, \Lambda)d\Lambda$ are the marginal likelihood and the marginal prior of $(\mathbf{X}_{pG}, p, G)$, respectively. Here we use $f(D|\cdots)$ to denote the sampling density of data $D$ given specified parameter values and $\Lambda$ to denote the mixture model parameters $(\epsilon, \mu, T)$. As $p$ or $G$ increases, the likelihood increases and it can be considered as a measure of fit. However, as $p$ or $G$ increases, the prior density decreases and it can be viewed as a penalty for more complex models.

Under equal prior probabilities for all $p$ in $p_{min} \leq p \leq p_{max}$ and for all $G$ in $G_{min} \leq G \leq G_{max}$,

$$\pi(\mathbf{X}_{pG}, p, G|D) \propto f(D|\mathbf{X}_{pG}, p, G)\pi(\mathbf{X}_{pG}|p, G).$$

Thus one only needs to compute the marginal likelihood and the marginal prior of $\mathbf{X}$ for each $p$ and $G$. Let $\pi(\mathbf{X}_{pG}|D) = \pi(\mathbf{X}_{pG}, p, G|D)$, $f(D|\mathbf{X}_{pG}) = f(D|\mathbf{X}_{pG}, p, G)$, and $\pi(\mathbf{X}_{pG}) = \pi(\mathbf{X}_{pG}|p, G)$ for notational simplicity. Oh and Raftery (2001) showed that $f(D|\mathbf{X}_{pG})$ is approximately proportional to $SSR_{pG}^{-m/2+1}$, where $SSR_{pG} = \sum_{i>j}(\delta_{ij}^{(pG)} - d_{ij})^2$ and $\delta_{ij}^{(pG)}$ is the Euclidean distance between the $\mathbf{x}_i$ and $\mathbf{x}_j$ of $\mathbf{X}_{pG}$.

Now consider computation of $\pi(\mathbf{X}_{pG})$. When $G = 1$, with $T = diag(t_1, \cdots, t_p)$ and independent $IG(\alpha, \beta_j)$ priors for $t_j$, $j = 1, ..., p$,

$$\pi(\mathbf{X}_{p1}) = (2\pi)^{-np/2}(\Gamma(\alpha + n/2)/\Gamma(\alpha))^p \prod_{j=1}^{p} \beta_j^{\alpha}(\beta_j + s_j/2)^{-(\alpha+n/2)}, \tag{12}$$

where $s_j$ is the $j$-th diagonal component of $nS_x = \sum(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$ (Oh and Raftery 2001).

When $G \geq 2$, the prior $\pi(\mathbf{X}_{pG})$ is not given in a closed form and needs to be estimated. From the relationship

$$\pi(\mathbf{X}_{pG}) = \frac{\pi(\mathbf{X}_{pG}|\Lambda^*)\pi(\Lambda^*)}{\pi(\Lambda^*|\mathbf{X}_{pG})} \tag{13}$$

for a fixed value $\Lambda^*$ of $\Lambda$, $\pi(\mathbf{X}_{pG})$ can be estimated from estimate of $\pi(\Lambda^*|\mathbf{X}_{pG})$. From Oh (1999) it can be shown that

$$\pi(\Lambda^*|\mathbf{X}_{pG}) = E[\pi(\varepsilon^*|K,G)\prod_{j=1}^{G}\pi(\mu_j^*|K,T_j^*,\mathbf{X}_{pG})\pi(T_j^*|K,\mu_j,\mathbf{X}_{pG})], \tag{14}$$

where the expectation is with respect to the joint distribution of $(K,\varepsilon,\mu,T)$ given $(\mathbf{X}_{pG},p,G)$. Since the conditional distributions of $\varepsilon,\mu_j,T_j$ are given in closed forms, $\pi(\Lambda^*|\mathbf{X}_{pG})$ can be easily estimated by using samples of $(K,\varepsilon,\mu,T)$ generated from MCMC algorithm. In theory any value of $\Lambda^*$ can be used but in practice $\Lambda^*$ close to the mode of $\Lambda$ seems to work well for efficiency point of view.

However, simple comparison of the posterior of $\mathbf{X}_{pG}$ can lead to the choice of large $p$ because of the shrinking effect. As described in Oh and Raftery (2001) there is a shrinking effect as the dimension $p$ increases, i.e., the scale (dispersion) of $\mathbf{X}$ tends to decrease as $p$ increases without altering the fit. This would yield larger $\pi(\mathbf{X}_{pG})$ for larger $p$ even when the likelihoods are the same and hence would favor larger $p$. To avoid this shrinking effect, Oh and Raftery (2001) have suggested comparing configurations in the same dimensional space. Specifically, they compare dimensions $p$ and $p-1$ through $\mathbf{X}_p$ and $\mathbf{X}_p^* = (\mathbf{X}_{p-1} : 0)$, a $n \times p$ matrix with the $p-1$ columns equals to $\mathbf{X}_{p-1}$ and the last column has all elements equal to 0. Note that $\mathbf{X}_p^*$ yields the same likelihood as $\mathbf{X}_{p-1}$ and it may be considered as an implantation of $\mathbf{X}_{p-1}$ in $p$-dimensional space.

When $G=1$, let $\mathbf{X}_p = \mathbf{X}_{p1}$. Since $f(D|\mathbf{X}_p^*) = f(D|\mathbf{X}_{p-1})$,

$$\begin{aligned}\frac{\pi(\mathbf{X}_p|D)}{\pi(\mathbf{X}_p^*|D)} &= \frac{f(D|\mathbf{X}_p)\pi(\mathbf{X}_p)}{f(D|\mathbf{X}_p^*)\pi(\mathbf{X}_p^*)}\\ &= [\frac{f(D|\mathbf{X}_p)}{f(D|\mathbf{X}_{p-1})}\frac{\pi(\mathbf{X}_p)}{\pi(\mathbf{X}_{p-1})}][\frac{\pi(\mathbf{X}_{p-1})}{\pi(\mathbf{X}_p^*)}]\\ &= [\frac{\pi(\mathbf{X}_p|D)}{\pi(\mathbf{X}_{p-1}|D)}][\frac{\pi(\mathbf{X}_{p-1})}{\pi(\mathbf{X}_p^*)}].\end{aligned}$$

Thus, $A_p = \pi(\mathbf{X}_{p-1})/\pi(\mathbf{X}_p^*)$ is a correction factor to the posterior density ratio of $\mathbf{X}_p$ and $\mathbf{X}_{p-1}$ for the shrinking effect. With $\alpha = 1/2$ and $\beta_j = s_j^{(p)}/2n$, an approximate unit information prior, for $t_j$, $A_p$ is given as

$$-2\ \log\ A_p = -2\ \log\frac{\pi(\mathbf{X}_{p-1})}{\pi(\mathbf{X}_p^*)} = H_n - n\log(\frac{s_p^{(p)}}{n}) + \sum_{j=1}^{p-1}\log(r_j^{(p)}) + (n+1)\frac{\log((n+1)}{(n+r_j^{(p)})}),$$

where $r_j^{(p)} = s_j^{(p)}/s_j^{(p-1)}$ and $H_n = -(n+1)\log(\pi) + 2\log(\Gamma((n+1)/2))$. The shrinking effect is related only to $p$ and not to $G$, so we use $A_p$ for all values of $G$ given the same $p$.

Now we propose a selection criterion, which we call MIC, as follows. Let

$$\begin{aligned} MIC_{1G} &= (m-2)\log SSR_{1G} - 2\log\pi(\mathbf{X}_{1G}) \\ MIC_{pG} &= \sum_{q=1}^{p} -2\log\frac{\pi(\mathbf{X}_{qG}|D)}{\pi(\mathbf{X}^*_{qG}|D)} \qquad (15) \\ &= \sum_{q=1}^{p} -2\log\frac{f(D|\mathbf{X}_{qG})}{f(D|\mathbf{X}_{q-1,G})}\frac{\pi(\mathbf{X}_{qG})}{\pi(\mathbf{X}_{q-1,G})} - 2\log A_q \qquad (16) \\ &= (m-2)\log(SSR_{pG}) - 2\log\pi(\mathbf{X}_{pG}) - 2\sum_{q=1}^{p}\log A_q. \qquad (17) \end{aligned}$$

Note that $(m-2)\log(SSR_{pG})$ can be considered as a measure of fit, $-2\log\pi(\mathbf{X}_{pG})$ plays the role of a penalty for complexity, and $-2\sum_{q=1}^{p}\log A_q$ is a cumulative correction factor for the shrinking effect. The values of $p$ and $G$ that yield the minimum of $MIC_{pG}$ are viewed as best.

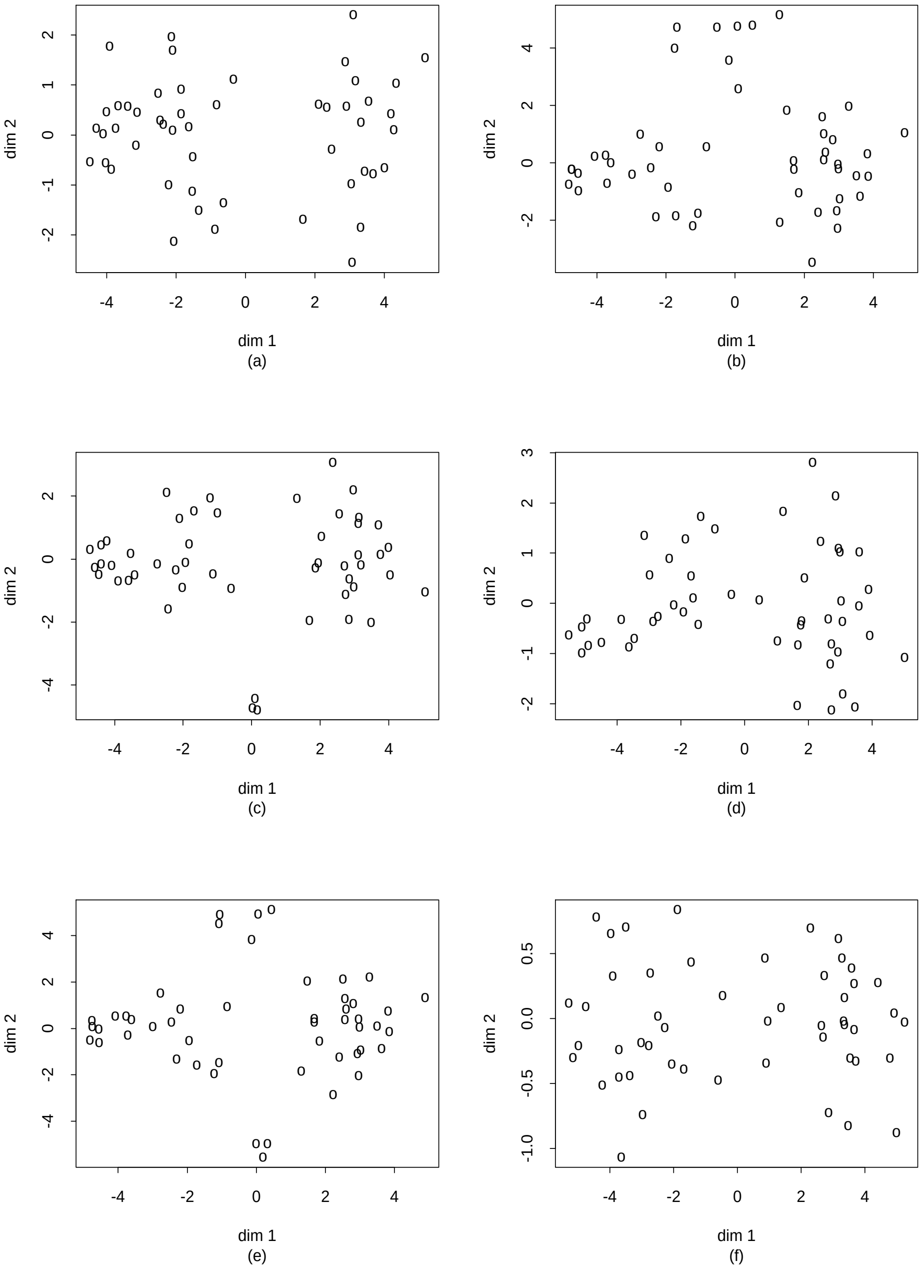
# 5 Examples

We apply the proposed method, which we call BMCD, and the model selection criterion MIC, to some simulated and real data sets.

In the simulation and Bank examples given in Sections 5.1 and 5.2, we use a general covariance structure for $T_j$, and for the Leukemia and Yeast examples in Sections 5.3 and 5.4 we use the same covariance structure for $T_j$, i.e., $T_1 = T_2 = \cdots = T_G$. In all the examples, we use $\bar{\mathbf{x}}$ as the prior mean of $\mu_j$ and we let $\alpha = p + 4$ and $B_j = (\alpha - p - 1)S_x$ for the hyper-parameters of the Inverted Wishart prior of $T_j$ in (3), where $\bar{\mathbf{x}}$ and $S_x$ are the average and sample covariance matrix of the initial $\mathbf{x}_i$'s. Thus, we use a common mean for all $\mu_j$ and a common vague prior for all $T_j$, and choose the scale parameter of the Inverted Wishart distribution so that the prior mean of $T_j$ is equal to $S_x$.

## 5.1 Simulation Examples

Six data sets with 50 objects each were generated from mixtures of bivariate normal distributions with various values of the mixture model parameters. Scatterplots of the true objects from the six data sets are given in Figure 1. In all cases the true dimension is $p = 2$ but the true numbers of clusters are different. The first set has two well-separated clusters and the second set has three well-separated clusters. The third set has two big clusters and a small cluster which may be considered as a group of outliers. The fourth set has two clusters with



(a)

(b)

(c)

(d)

(e)

(f)

Figure 1: Scatterplots of true objects in the simulation data. There are two well-separated clusters in (a), three well-separated clusters in (b), two big clusters and a small cluster of outliers in (c), two clusters with different covariance structures in (d),two big clusters and two small clusters in a symmetric position in (e), and two close clusters in (f).

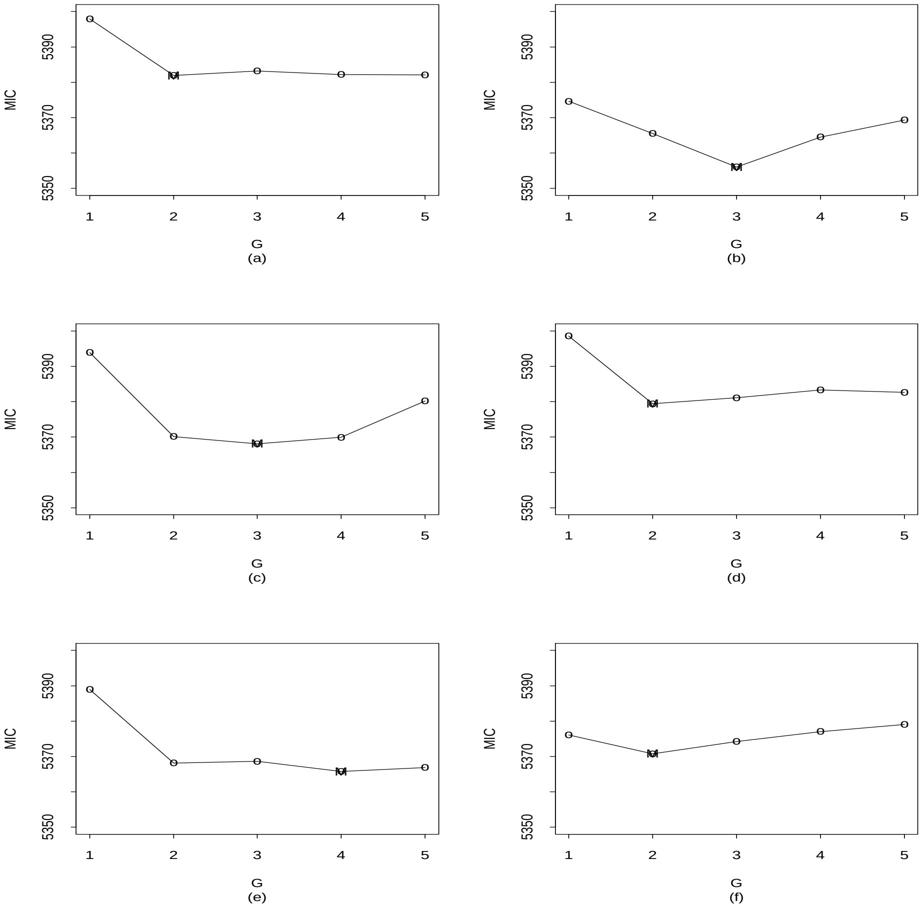
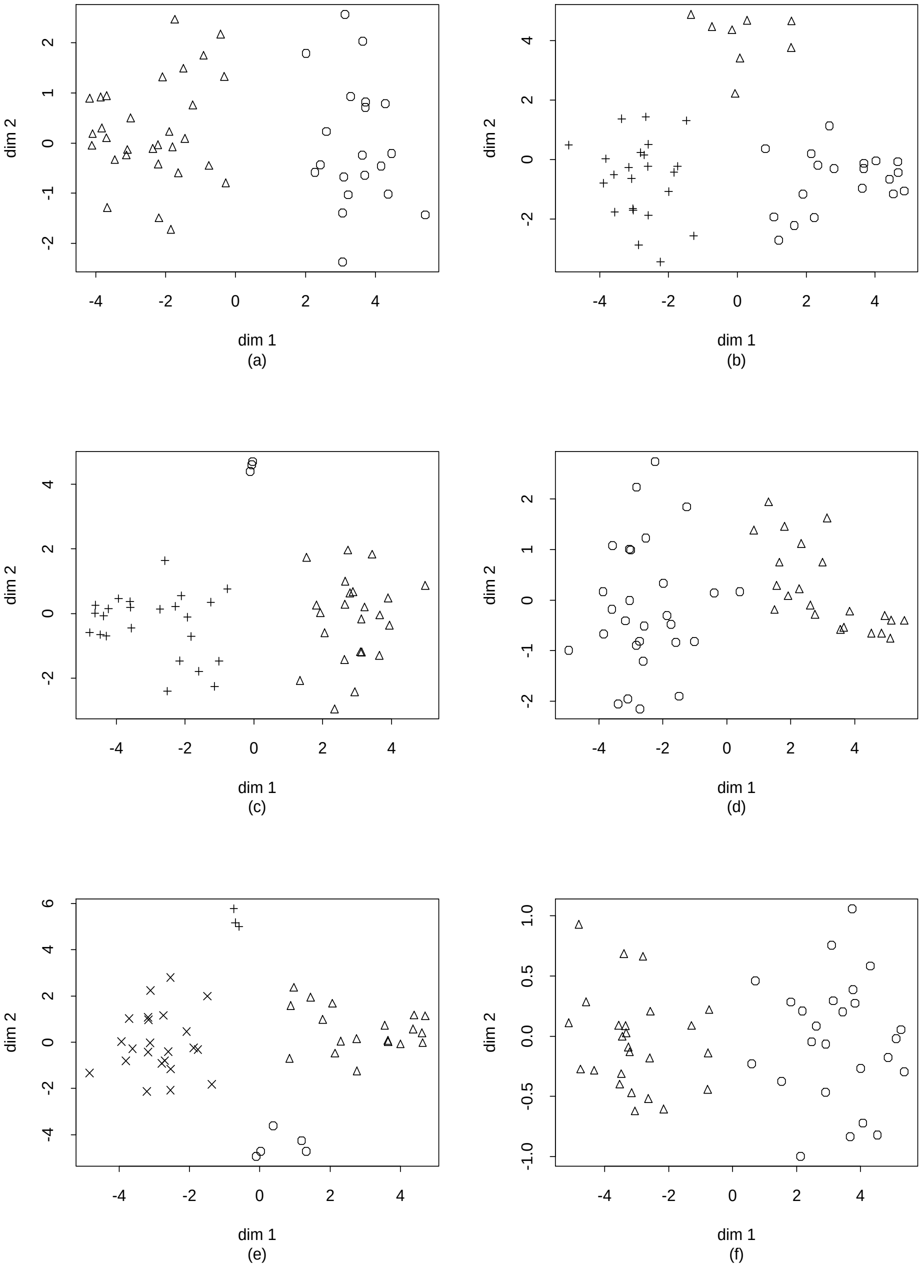


Figure 2: Plots of MIC for $p = 2$ in the six simulation data sets. In all cases MIC chooses the correct number of clusters.



(a) (b) (c) (d) (e) (f)

Figure 3: Scatterplots of the estimated object configuration and the classification from BMCD with the optimal $p$ and $G$ in the six simulation data sets. Different symbols represent different clusters.

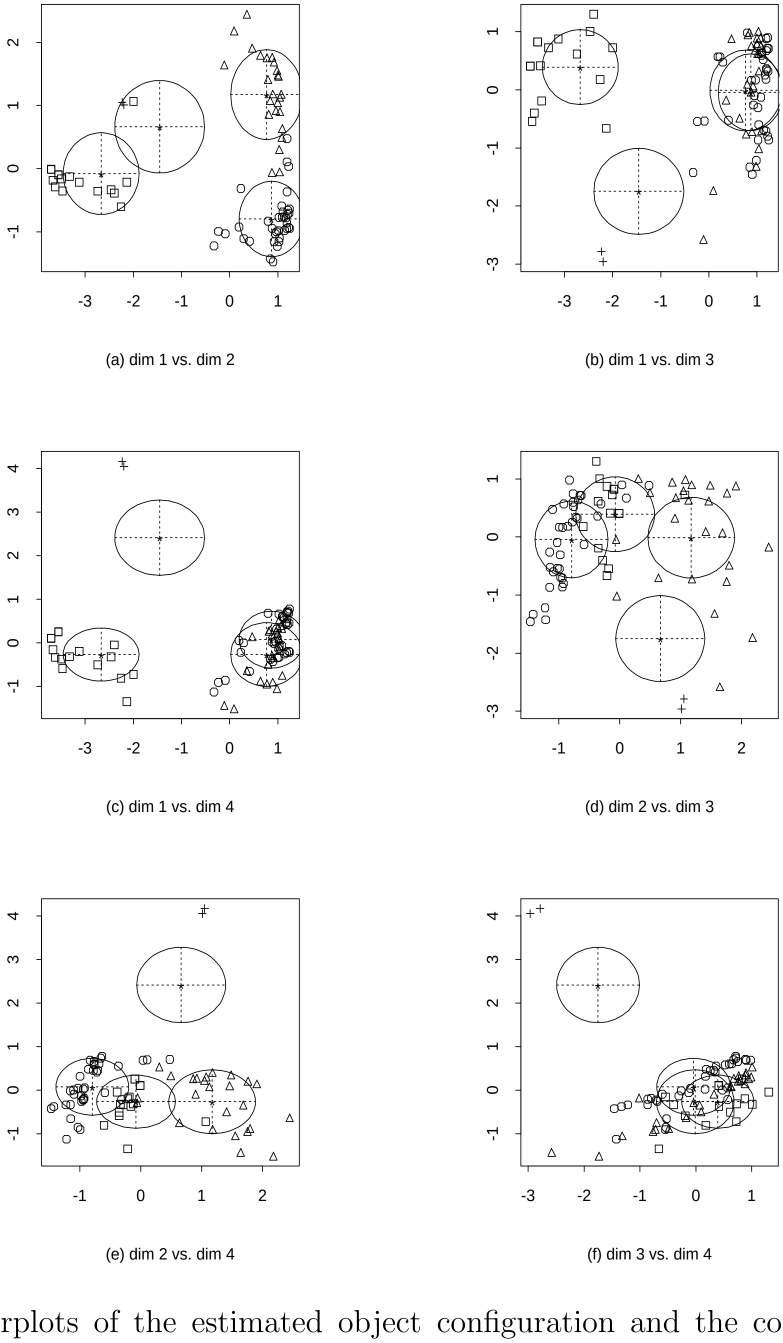
different covariance structures. The fifth set has two big clusters and two small clusters that are symmetrically located. The last set has two close clusters. For each data set, Euclidean distances $\delta_{ij}$ between pairs of objects were computed and a $50 \times 50$ matrix of observed dissimilarity measures $d_{ij}$ was obtained by generating the $d_{ij}$ from a normal distribution with mean $\delta_{ij}$ and standard deviation 0.3, restricted to be positive.

For each data set, we applied BMCD with 20,000 MCMC iterations, of which 5,000 were discarded for burn-in, and we computed MIC for various values of $p$ and $G$. In all cases, the MIC values for $p$ other than 2 were much larger than the MIC value for $p = 2$ for every value of $G$ considered, so MIC correctly identifed the true dimension $p = 2$. Figure 2 presents plots of MIC for various $G$ when $p = 2$ for the six data sets. In all cases, MIC selected the correct number of clusters, though MIC's preference for the selected $G$ was not as strong as for $p$. Also, the estimated object configurations from BMCD were good, as can be seen in Figure 3.

## 5.2 Lloyds Bank Data

We now consider dissimilarity measures between the careers of 80 employees of Lloyds Bank in England during the period 1905–1950 (Stovel et al. 1996), computed using the optimal alignment method of Sankoff and Kruskal (1983), as applied to career data by Abbott and Hrycak (1990). Note that the dissimilarity measures in this data set are not Euclidean distances and may not satisfy some properties of typical metric distances. Oh and Raftery (2001) analyzed the data and MIC chose $p = 8$ as the optimal dimension. After removing two outlying employees who had extremely short careers at the bank, MCLUST was applied to the estimate of $\mathbf{X}$ from BMDS. It chose $G = 3$ and yielded a reasonable classification of objects. The first group consisted of 16 employees who had short careers at the bank and spent all or most of their careers at the lowest clerk rank, and the second group consisted of 30 employees who had long careers at the bank and spent all or most of their careers at the lowest clerk rank. The third group consisted of 32 employees, 24 of whom were promoted to managers and 8 of whom had medium length careers and ended at the clerk level.

We applied BMCD to the data with 40,000 MCMC iterations, of which the first 10,000 were discarded as burn-in. MIC chose $p = 8$ and $G = 4$, which coincides with the results from the previous analysis. The first group was identical to the first group found in the previous analysis. The second and the third groups were almost identical to those found in the previous study except that the 8 employees who had medium length careers and ended at the clerk level were clustered together with those who had long careers and ended at the



(a) dim 1 vs. dim 2

(b) dim 1 vs. dim 3

(c) dim 1 vs. dim 4

(d) dim 2 vs. dim 3

(e) dim 2 vs. dim 4

(f) dim 3 vs. dim 4

Figure 4: Scatterplots of the estimated object configuration and the component density functions from BMCD for the Lloyds Bank data.

clerk level in BMCD, rather than with the managers. This is more satisfactory, substantively. The fourth group consisted of the two outliers which were removed before clustering in the previous analysis. Thus, BMCD picked up the outliers during the process and yielded a more sensible clustering than the previous analysis. Figure 4 shows pairwise scatter plots of object configurations and the estimated component densities for the first four coordinates. Note that in Figure 4, the component density for the outliers has mean close to zero and a large covariance since there are only two objects in the group and the effect of the prior is dominant for this component. From Figure 4, it seems that BMCD gives clear separation between clusters and takes care of the outliers.

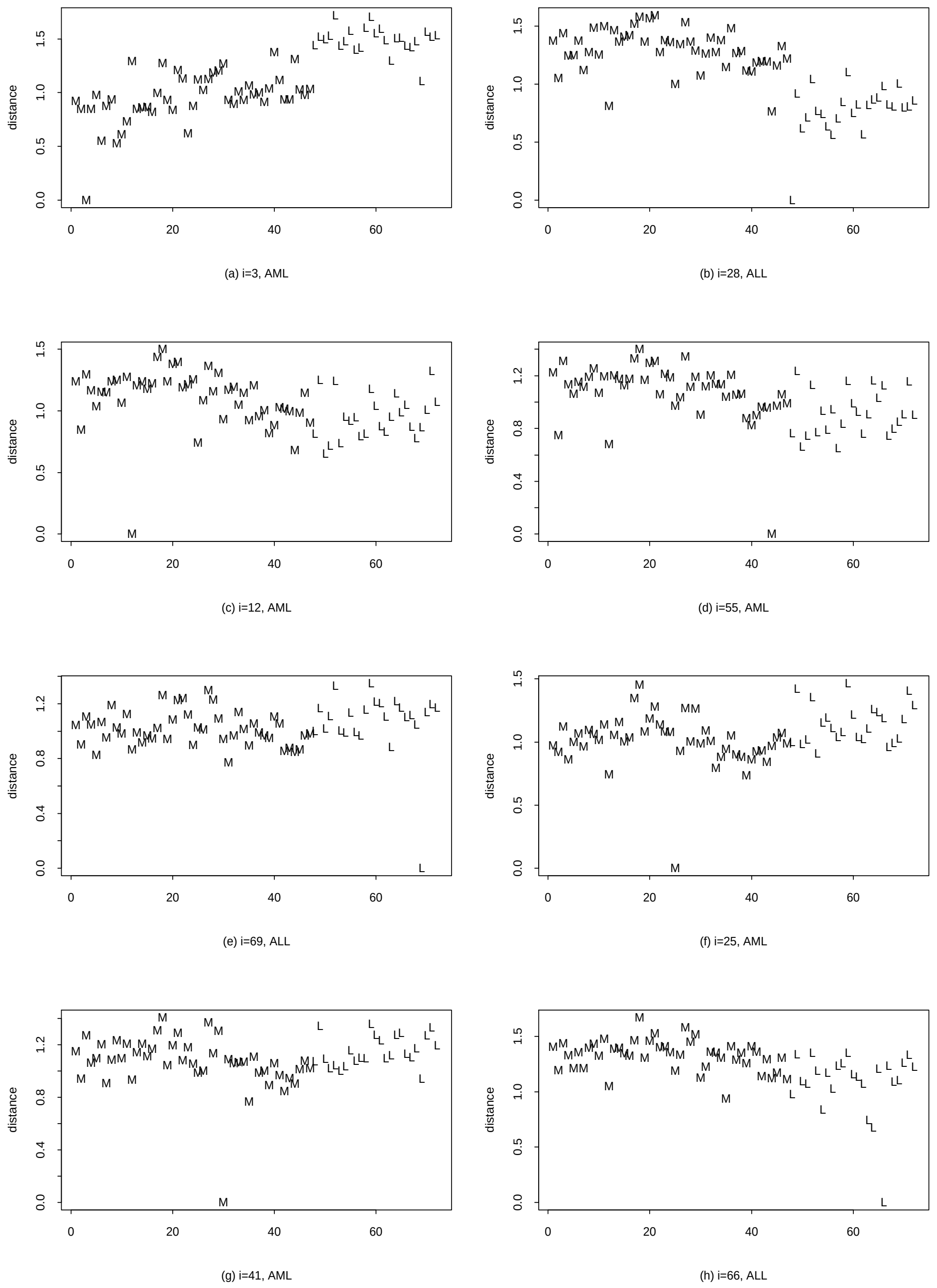
### 5.3 Leukemia Data

Golub et al. (1999) used gene expression data on 50 genes and 72 acute leukemia patients to classify the patients into different types of leukemia. The 50 genes are believed to be informative about the distinction between acute myeloid leukema (AML) and acute lymphoblastic leukemia (ALL) in the known samples.

We follow the standardization process given by Getz et al. (2000) and compute the Euclidean distance between genes, yielding a dissimilarity matrix similar to a correlation matrix. Due to the standardization, the mean of each gene is set to zero and this reduces the true dimension of the objects from 50 to 49.

Investigation of plots of dissimilarity measures between pairs of individuals shows that there are two big groups and most dissimilarity measures between individuals in the same group are small while those between individuals in different groups are large. Figure 5 (a) and (b) are typical plots for individuals in the first and the second groups, respectively. However, there are some individuals, such as numbers 12 and 55, who seem to be misclassified since they have smaller dissimilarities with those in the different group and larger dissimilarities with those in the same group as shown in Figure 5 (c)-(d). Also there are a few individuals whose distances do not clearly show their closeness to either group as shown in Figure 5 (e)-(h).

In all the examples we analyzed, MIC showed a sharp drop at the same optimal value of $p$ for all values of $G$, so that the choice of $G$ does not affect the choice of dimension. This is because the object configurations are not much different for different $G$s when the same $p$ is used. Thus, one may choose optimal dimension $p$ with $G = 1$ and then choose optimal $G$ with the selected $p$. In other words, one may apply BMDS to choose the dimension and then apply BMCD with the optimal dimension. This would reduce computation time significantly.



(a) i=3, AML

(b) i=28, ALL

(c) i=12, AML

(d) i=55, AML

(e) i=69, ALL

(f) i=25, AML

(g) i=41, AML

(h) i=66, ALL

Figure 5: Distances for some selected individuals in the Leukemia data (the id numbers and the type of Leukemia are given at the bottom of each figure). Typical distances for individuals in AML and ALL groups are shown in (a) and (b), respectively. Figures (c)-(d) presents distances for individuals who have smaller distances with those in the different group and larger distances with those in the same group. Figures (e)-(h) presents distances for individuals whose distances do not clearly show their closeness to either group.

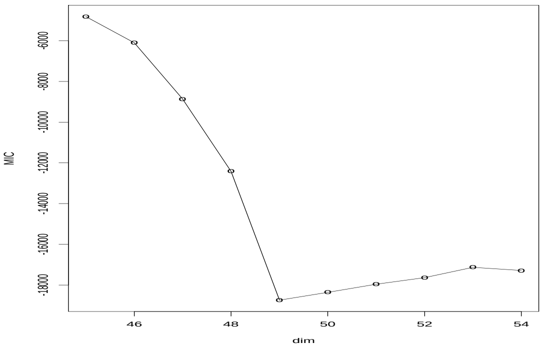


Figure 6: MICs for Leukemia data with $G = 1$.

Note that although $p$ and $G$ are chosen sequentially, the estimation of object configuraton and clustering are still done simultaneously.

Following the above suggestion, BMDS (i.e. BMCD with $G = 1$) was applied to the data and MDSIC clearly chose $p = 49$, which is the true dimension, as shown in Figure 6. We used 18,000 MCMC iterations, of which 3,000 were discarded as burn-in.

We next applied the method with $p = 49$ for various values of $G$ and MIC values are presented in Table 1. MIC clearly chooses $G = 2$, which is the correct number of groups for this dataset. Three individuals, numbers 12, 55 and 69, are misclassified, so the misclassification rate is 4.2%. The 41st individual is classified into the correct AML group, but his or her posterior membership probability for AML is only 0.51, showing large uncertainty. This result makes sense in light of the dissimilarity measures in Figure 5.

For visual display of the clusters and objects in low dimensional space, we applied the method with $p = 2, 3$ for various values of $G$ and the results are presented in Table 1. MIC chooses $G = 2$ groups both when $p = 2$ and when $p = 3$. Classification results from $p = 2, 3$ are the same as those from $p = 49$ with $G = 2$ except for the 41st and the 69th individuals. When $p = 2$ and 3, the 41st individual is misclassified into the ALL group but with membership probability of only 0.51, and the 69th individual is correctly classified into the AML group when $p = 2, 3$. However, the membership probability for the 69th individual is about 0.56 when $p = 2, 3$ while it is close to 1.0 when $p = 49$, showing significant uncertainty when $p = 2, 3$. Thus, in terms of clustering, 2 or 3 dimensions from BMCD does as well as the true 49 dimensions for all the individuals except for the 69th.

Figure 7 shows estimated object configurations with their classifications when $p = 3$ and $G = 2$. It is interesting to observe that clusters can be very well identified in the plot of the

Table 1: MIC values for Leukemia data for $p = 2, 3, 49$

| G | | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| $p = 2$ | MIC | 14444 | 14425 | 14428 | 14440 |
| $p = 3$ | MIC | 12986 | 12970 | 12973 | 12979 |
| $p = 49$ | MIC | -11058 | -22457 | -22432 | -22320 |

first two coordinates from BMCD. And it is hard to see clusters in the plots of the other coordinates. We have observed this in most of data sets we analyzed, so in some cases only a few coordinates from BMCD would do as well as all the coordinates, in terms of clustering.

To assess the benefits from simultaneous rather than separate estimation of object configuration and clustering, we compared BMCD with a two-stage scheme, which estimates object configuration and then applies model-based clustering, when $p = 3$ and $G = 2$. For a fair comparison, we used the same priors, and we applied the same MCMC procedures for posterior estimation of the parameters and use the estimated object configuration from BMCD as input data in the model-based clustering of the two-stage scheme. Note that the only difference between the two methods lies in the randomness of object configuration. In most cases the estimated membership probabilities were more extreme in the two-stage method. This may be because it is more likely to assign each object to a certain group when $\mathbf{X}$ is fixed than when it is random. More extreme probabilities yield smaller posterior standard deviations for the probabilities, suggesting that sequential application of MDS and model-based clustering can underestimate the clustering uncertainties.

## 5.4 Yeast Cell Cycle Data

The Yeast cell cycle data (Cho et al. 1998) consist of gene expression levels of approximately 6000 genes over 17 time points. Yeung et al. (2001) used a subset of this data consisting of 384 genes whose expression levels peak at different time points corresponding to five known phases of the cell cycle (Cho et al. 1998).

We normalized the data analyzed by Yeung et al. (2001) using a typical standardization method, subtracting the mean and dividing by the standard deviation, for each gene. We then computed Euclidean distance for each pair of genes and used the Euclidean distances as dissimilarity measures. Due to the normalization, we lose one dimension and hence the true dimension of object configuration from the dissimilarity matrix is 16.

Yeung et al. (2001) applied the MCLUST model-based clustering software (Fraley and

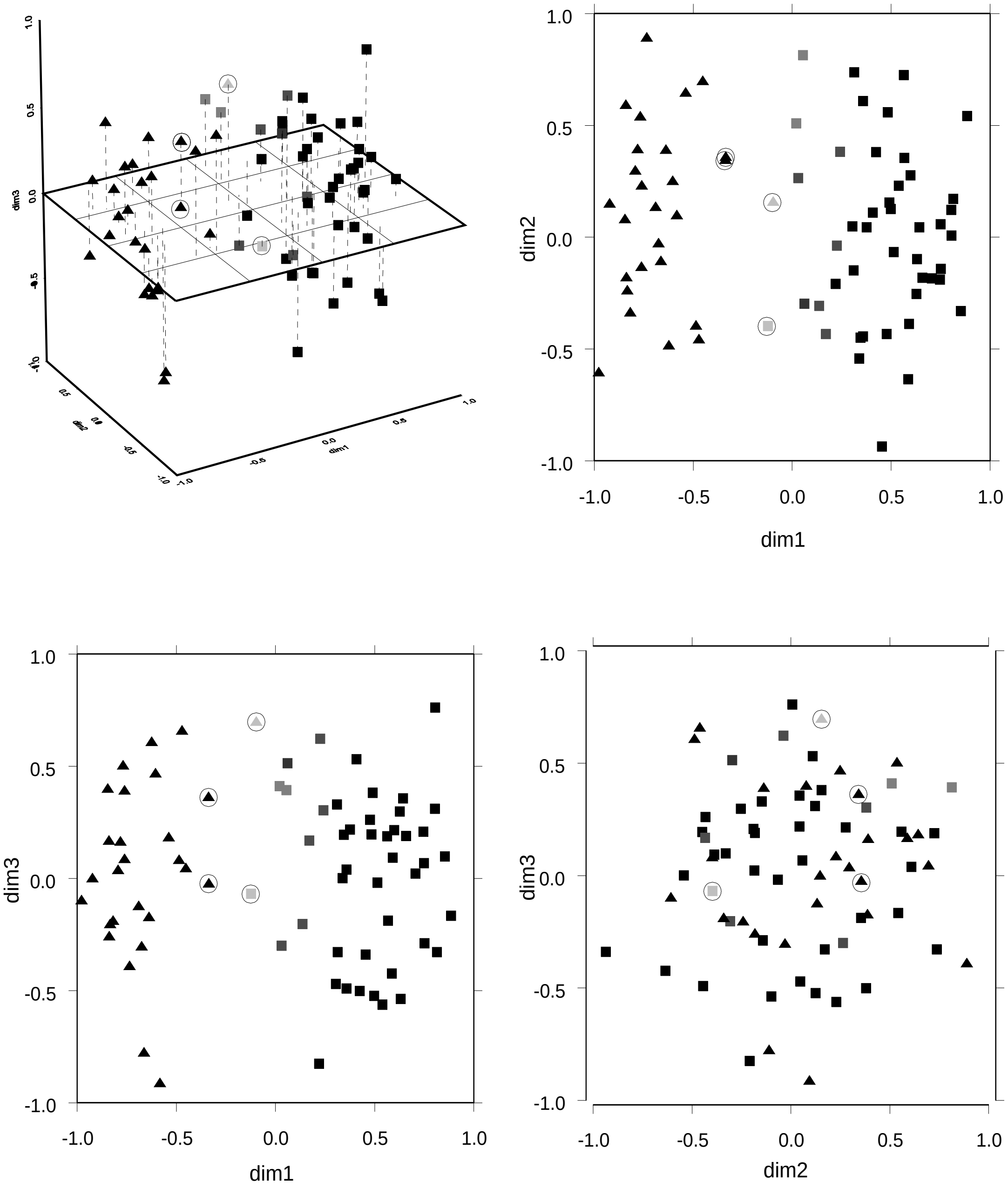


Figure 7: Three-dimensional and pairwise scatter plots for the object configuration and their classification from BMCD for Leukemia data. The sizes of membership probabilities are represented by the darkness of symbols, (black for probability 1 and gray for probability 0.5). Misclassified objects are marked by circles.

Table 2: MIC values for Yeast data for $p = 2, 3, 16$.

| G | | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| $p = 2$ | MIC | 844334 | 844232 | 844207 | 844185 | 844173 | 844201 |
| $p = 3$ | MIC | 772569 | 772459 | 772402 | 772423 | 772407 | 772396 |
| $p = 16$ | MIC | -1346244 | -1346602 | -1346788 | -1346521 | -1347010 | -1346648 |

Raftery 1999) to the 384 × 17 dataset and showed that the "EEE" model, which assumes that the mixture components have the same covariance matrix, gives a better fit than other models, so we made this assumption in applying BMCD to this data set. We used 18,000 MCMC iterations, of which we discarded 3,000 as burn-in.

First, to choose the optimal dimension of objects, BMDS was applied for $p = 1$ to 18. It indicated clear evidence for $p = 16$ which is the correct dimension in this example. Next BMCD was applied with $p = 16$ for $G = 3$ to 8. Values of MIC are given in Table 2. MIC reaches a minimum at $G = 7$ but has a first local minimum at $G = 5$, and here we consider $G = 5$ and $G = 7$ as possible optimal numbers of groups.

We applied BMCD with $p = 2, 3$ for visualization and parsimony. Values of MIC are given in Table 2. When $p = 2$, MIC chooses $G = 7$ and when $p = 3$ it chooses $G = 8$, but MIC has about the same value at $G = 5$ and $G = 7$. Figure 8(a) presents the two-dimensional object configuration from BMCD with the actual five known phases. There are significant overlaps between the actual clusters. Figure 8(b) shows the estimated object configuraton and classification results from BMCD with $p = 2$ and $G = 5$. It can be seen that BMCD yields a reasonable clustering of the objects.

We compared the clustering results for $p = 2$, $p = 3$, and $p = 16$, when $G = 5$. There are 20 mismatches between $p = 2$ and $p = 16$, and 22 mismatches between $p = 3$ and $p = 16$. Thus, the proportion of mismatches is less than 6% between the low dimensional clustering and the clustering with the true dimension. Many of the mismatched genes show significant clustering uncertainties in low dimensions, suggesting that the assessment of uncertainty is accurate in these cases. We next computed the proportion of mismatches between clustering from BMCD and the actual five clusters. These are 0.268, 0.266 and 0.279 for $p = 16$, $p = 3$, and $p = 2$, respectively, indicating that there is almost no difference in clustering quality between the different dimensions.

As for the Leukemia data, we performed the two-stage scheme of MDS plus model-based clustering with $p = 2$ and $G = 5$, and compared its membership probabilities with those from BMCD. The main conceptual difference between the two methods lies in whether $\mathbf{X}$ is fixed or randomly generated at each MCMC iteration. In almost all cases the two-stage scheme provides more extreme membership probabilities, resulting in a smaller posterior standard deviation. This again suggests that first estimating object configuration and then clustering (as opposed to doing both simultaneously as in BMCD) does not take into account the variation in object configuration when clustering and that it may underestimate the clustering uncertainties.

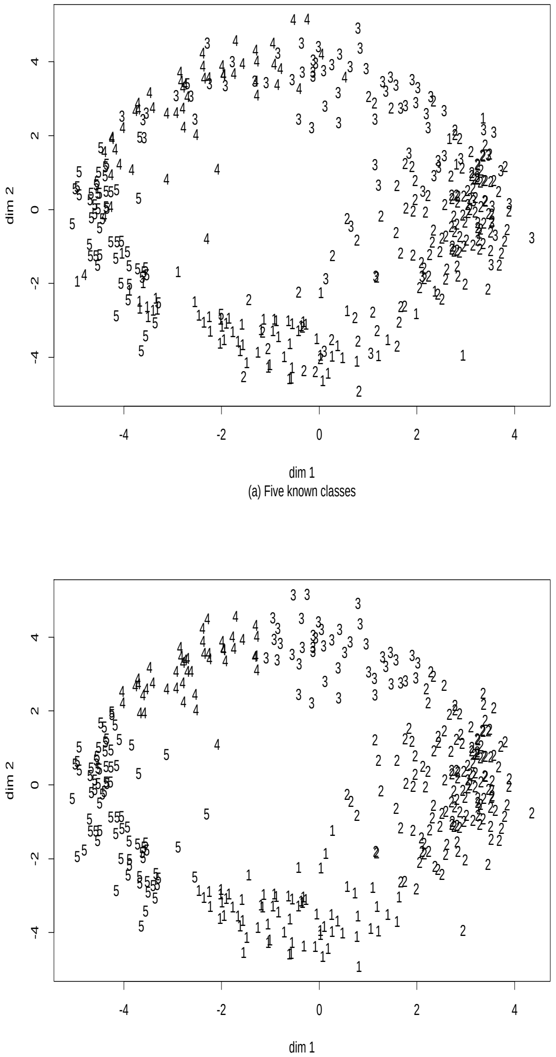
# 6 Discussion

We have proposed a model-based clustering method for the situation where the data consist of dissimilarity measures between pairs of objects. It is also useful for clustering objects in low dimensional space for visual display and parsimony even when the object coordinates are given, but are high-dimensional.

Hierarchical models are used to represent the possible sources of error, namely measurement errors in the dissimilarities, errors in estimating object configuration, and errors in clustering the objects. A probabilistic model is used for the observed dissimilarities and a mixture model is used for the unobserved latent object configuration. The object configuration, the mixture model parameters, and the objects' group memberships are estimated simultaneously via Bayesian inference using MCMC. The object configuration can be used for display of objects and the mixture parameters can be used for clustering objects. Thus, the method performs MDS and model-based clustering simultaneously, taking account of the errors simultaneously rather than sequentially and hence yielding a reasonable measure of clustering uncertainty.

In contrast, other methods of clustering from dissimilarity measures either do not incorporate all the errors or do not take them into account simultaneously. In Table 3 we summarize some possible alternative types of technique for clustering using dissimilarity data. The first consists of heuristic clustering methods which can cluster directly from dissimilarities, such as hierarchical agglomerative clustering, $k$ means, and self-organizing maps. The second scheme is a sequential application of a typical MDS, which gives object configuration without estimation error, and model-based clustering, which provides clustering uncertainty. The third scheme is a sequential application of BMDS, which provides object configuration as well as estimation error, and model-based clustering.

When the estimated dimension is high, we have compared BMDS with the selected di-



(a) Five known classes

(b) Classification from BMCD

Figure 8: Scatterplots of object configuration from BMCD with $p = 2$ in Yeast data and their classifications: (a) presents the actual five clusters and (b) presents the classification from BMCD with $G = 5$.

Table 3: Three Sources of Errors in Clustering with Dissimilarities

| Error Sources | Heuristic clustering | MDS+ MBC | BMDS + MBC | BMCD |
|---|---|---|---|---|
| Dissimilarity | no | no | yes | yes |
| Object configuration | not applicable | no | yes | yes |
| Clustering | no | yes | yes | yes |
| Simultaneous consideration | no | no | no | yes |

mension with BMDS with low dimension (2 or 3). We found that the clustering results were very similar, and that those misclassified in the low-dimensional analysis had high clustering uncertainties, which is good. Thus, in practice BMDS low dimensional configurations may be good enough for many purposes, especially if it is followed up with more intensive investigation of objects with high clustering uncertainty.

We have proposed a Bayesian criterion, MIC, for simultaneously selecting the object dimension and the number of clusters, which is easy to compute from MCMC output. In our simulations and in real examples, it worked reasonably well in all cases. MIC varied more between dimensions than between numbers of clusters, and the choice of dimension was not affected by the choice of the number of clusters. Thus, as an approximation we suggest selecting the dimension assuming one cluster (i.e. using BMDS), and then choosing the number of clusters given the selected dimension. This greatly reduces computation time.

One important area where data come in the form of measures on pairs of objects is social networks, where data consist of the presence or absence (or in some cases the intensity) of ties between actors. Hoff, Raftery and Handcock (2002) used ideas similar to those of Oh and Raftery (2001) to represent actors in a social network by positions in a Euclidean latent space and estimate the positions. The model used was the same as that of Oh and Raftery (2001), except that the conditional distribution of "dissimilarities" (in the social network case, presence or absence of ties) given distances was taken to be binary with a conditional probability specified by logistic regression, rather than truncated normal. The analysis of social network data is often motivated by questions about the presence and nature of clusters in the network, and these are often answered fairly heuristically. It would seem straightforward to extend the present approach to social network data, again modeling presence or absence of ties as conditionally binary with a probability depending on distance in a logistic regression manner. This could provide a more formal way of answering questions about clustering in social networks.

# References

[1] Abbott, A. and Hrycak, A. (1990), "Measuring Sequence Resemblance," *American Journal of Sociology*, 96, 144–185.

[2] Banfield, J.D. and Raftery, A.E. (1993). "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics*, 49, 803–821.

[3] Bishop, C. (1995), *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford.

[4] Borg, I. and Groenen, P. (1997). *Modern Multidimensional Scaling*, Springer-Verlag, New York, Berlin.

[5] Buttenfield, B. and Reitsma, R.F. (2002), "Loglinear and Multidimensional Scaling Models of Digital Library Navigation", *Internation Journal of Human-Computer Studies*, 57, 101-119.

[6] Celeux, G., Hurn, M., and Robert C.P. (2000) "Computational and Inferential Difficulties with Mixture Posterior Distribution", *Journal of the American Statistical Association*, 95, 957-970.

[7] Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D.J., Lockhart, D.J., and Davis, R.W. (1998), "A Genome-wide Transcriptional Analysis of the Mitotic Cell Cycle", *Molecular Cell*, 2, 65-73.

[8] Condon, E., Golden, B., Lele, S., Raghavan, S., and Wasil, E. (2002), "A Visualization Model Based on Adjacency Data", *Decision Support Systems*, 33, 349-362.

[9] Courrieu, P. (2001), "Two Methods for Encoding Clusters", *Neural Networks*, 14, 175-183.

[10] Elvevag, B. and Storms, G. (2002), "Scaling and Clustering in the Study of Semantic Disruptions in Patients with Schizophrenia: a Re-evaluation", *Schizophrenia Research*, in press.

[11] Fraley, C. and Raftery, A.E. (1999), "MCLUST: Software for Model-based Cluster Analysis", *Journal of Classification*, 16, 297-306.

[12] Fraley, C. and Raftery, A.E. (2002), "Model-Based Clustering, Discriminant Analysis, and Density Estimation", *Journal of the American Statistical Association*, 97, 611–631.

[13] Getz, G., Levine, E., and Domany, E. (2000), "Coupled Two-way Clutering Analysis of Gene Microarray Data", *Proceedings of National Academy of Sciences*, 97, 12079-12084.

[14] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, M.L., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., and Lander, E.S. (1999), "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring", *Science*, 286, 531-537.

[15] Hastings, W.K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications", *Biometrika*, 57, 97-109.

[16] Hedenfalk, I.A., Ringer, M., Trent, J., Borg, A. (2002). "Gene Expression in Inherited Breast", *Cancer Research*, 84, 1-34.

[17] Hoff, P., Raftery, A.E. and Handcock, M. (2002), "Latent Space Approaches to Social Network Analysis," *Journal of the American Statistical Association*, 97, 1090–1098.

[18] Kohonen, T. (2001), *Self-Organizing Maps*, Springer-Verlag, New York.

[19] MacQueen, J. (1967), "Some Methods for Classification and Analysis of Multivariate Observations," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, eds. L.M.LeCam and J. Neyman, Berkeley, Calif.: University of California Press, pp. 281–297.

[20] McLachlan G., and Peel, D. (2000), *Finite Mixture Models*, Wiley, New York.

[21] Nikkila, J., Toronen, P., Kaski, S., Venna, J., Castren, E., and Wong, G. (2002), "Analysis and Visualization of Gene Expression Data Using Self-Organizing Maps, *Neural Networks*, 15, 953-966.

[22] Oh, M-S. (1999), "Estimation of Posterior Density Functions from a Posterior Sample", *Computational Statistics & Data Analysis*, 29, 411-427.

[23] Oh, M-S. and Raftery, A. (2001), "Bayesian Multidimensional Scaling and Choice of Dimension", *Journal of the American Statistical Association*, 28, 259-271.

[24] Priem, R.L., Love, L., and Shaffer, M.A. (2002), "Executives' Perceptions of Uncertainty Sources: A Numerical Taxonomy and Underlying Dimensions", *Journal of Management*, 28, 725-746.

[25] Ren, S. and Frymier, P.D. (2003), "Use of Multidimensional Scaling in the Selection of Wastewater Toxicity Test Battery Components", *Water Research*, 37, 1655-1661.

[26] Sankoff, D. and Kruskal, J.B. (1983), *Time Warps, String Edits and Macromolecules,* Reading, Mass.: Addison-Wesley.

[27] Schutze, H. and C. Silverstein (1997), "Projections for Efficient Document Clustering", *ACM SIGIR 97*, 74-81.

[28] Sneath, P.H.A. (1957), "The Application of Computers to Taxonomy," *Journal of General Microbiology*, 17, 201–206.

[29] Sokal, R.R. and Michener, C.D. (1958), "A Statistical Method for Evaluating Systematic Relationships," *University of Kansas Scientific Bulletin*, 38, 1409–1438.

[30] Stephens, M. (2000). "Dealing with Label-Switching in Mixture Models," *Journal of the Royal Statistical Society, Series B*, 62, 795–809.

[31] Stovel, K., Savage, M., and Bearman, P. (1996), "Ascription into Achievement: Models of Career Systems at Lloyds Bank, 1890-1970", *American Journal of Sociology*, 102, 358-399.

[32] Tibshirani, R., Lazzeroni, L., Hastie, T., Olshen, A., and Cox, D. (1999), "The Global Pairwise Approach to Radiation Hybrid Mapping", Technical Report, Department of Statistics, Stanford University.

[33] Welchew, D.E., Honey, G.D., Sharma, T., Robins, T.W., and Bullmore, E.T. (2002), "Multidimensional Scaling of Integrated Neurocognitive Function and Schizophrenia as a Disconnexion Disorder", *NeuroImage*, 17, 1227-1239.

[34] Yeung, K., Fraley, C., Murua, A., Raftery, A.E., and Ruzzo, W.L. (2001), "Model-based Clustering and Data Transformation for Gene Expression Data", *Bioinformatics*, 17, 977-987.

[35] Yin, H. (2002), "Data Visualization and Manifold Mapping Using the ViSOM", *Neural Networks*, 15, 1005-1016.

[36] Young, F.W. (1987), *Multidimensional scaling, History, Theory, and Applications*, , edited by Hammer, R.M., Lawrence Erlbaum Associates, Publishers, Hillsdale, New Jersey.

## APPENDIX

### A. Procrustean transformation

- Step 0: Let $\mathbf{J}$ be the centering matrix, i.e., $\mathbf{J} = \mathbf{I} - 1/n\mathbf{1}\mathbf{1}'$, where $\mathbf{I}$ is the identity matrix and $\mathbf{1}$ is the vector of all 1's.
- Step 1: Compute $\mathbf{C} = \mathbf{X}^{*\prime}\mathbf{J}\mathbf{X}$.
- Step 2: Compute the singular value decomposition of $\mathbf{C}$, i.e., $\mathbf{C} = PDQ'$, where $P$ and $Q$ are orthogonal matrices and $D$ is a diagonal matrix.
- Step 3: Let $\mathbf{T} = QP'$.
- Step 4: Let $\mathbf{t} = 1/n(\mathbf{X}^* - \mathbf{X}\mathbf{T})'\mathbf{1}$.
- Step 5: Transform $\mathbf{X}$ by $\mathbf{X} = \mathbf{X}\mathbf{T} + \mathbf{1}\mathbf{t}'$.

### B. Relabeling Procedure

Let $\boldsymbol{\theta}$ be a $d$ dimensional vector of all the parameters in the mixture distribution, and $J$ be the number of components in the mixture.

- Step 0. Estimate elements of $\boldsymbol{\theta}$ and their variances using samples taken before the first label switching. Let $\boldsymbol{\theta}^0$ and $\mathbf{s}^0$ be the above estimates. Permute the labeling of the latent classes and deduce $\boldsymbol{\theta}^1, .., \boldsymbol{\theta}^{J!-1}$ and $\mathbf{s}^1, .., \mathbf{s}^{J!-1}$.
- Step 1. For each sample of $\boldsymbol{\theta}$, do :

  (1) Get $l^*$ which minimizes the squared-distances

  $$||\boldsymbol{\theta} - \boldsymbol{\theta}^l||^2 = \sum_{i=1}^{d} \frac{(\theta_i - \theta_i^l)^2}{s_i},$$

  for $l = 0, ..., J! - 1$, where $\theta_i$ and $s_i$ are the $i-$th coordinate of $\boldsymbol{\theta}$ and $\mathbf{s}$, respectively. Allocate $\boldsymbol{\theta}$ to the label given by $l^*$. Switch permutations $l^*$ and 0.

  (2) Update $\boldsymbol{\theta}^0$ and $\mathbf{s}^0$ and derive $\boldsymbol{\theta}^1, .., \boldsymbol{\theta}^{J!-1}$ and $\mathbf{s}^1, .., \mathbf{s}^{J!-1}$ by permutation.